AUDITA: A New Dataset to Audit Humans vs. AI Skill at Audio QA

Tasnim Kabir
University of Maryland
tkabir1@umd.edu

Dmytro Kurdydyk
Davidson College
dmkurdydyk@davidson.edu

Aadi Palnitkar
University of Maryland
apalnitk@terpmail.umd.edu

Liam Dorn
Columbia University
lmd2243@columbia.edu

Ahmed Haj Ahmed
Haverford College
ahajahmed@haverford.edu

Jordan Lee Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Abstract

Existing audio question answering benchmarks largely emphasize sound event classification or caption-grounded queries, often enabling models to succeed through shortcut strategies, short-duration cues, lexical priors, dataset-specific biases, or even bypassing audio via metadata and captions rather than genuine reasoning Thus, we present AUDITA (**A**udio **U**nderstanding from **D**iverse **I**nternet **T**rivia **A**uthors), a large-scale, real-world benchmark to rigorously evaluate audio reasoning beyond surface-level acoustic recognition. AUDITA[1] comprises carefully curated, human-authored trivia questions grounded in real-world audio, designed to stress robust auditory reasoning through challenging distractors and long-range temporal dependencies, using probing queries that cannot be answered from isolated text or sound cues alone. Human average accuracy of 32.13% shows both the challenge of the task while demonstrating meaningful comprehension of the audio. In stark contrast, state-of-the-art audio question answering models perform poorly, with average accuracy below 8.86%. Beyond raw accuracy, we apply Item Response Theory (IRT) to estimate latent proficiency, question difficulty, and expose systematic deficiencies of the models and data.

1 Introduction

Question answering (QA) is a central paradigm for evaluating language understanding, with modern benchmarks such as Natural Questions (Kwiatkowski et al., 2019) and large-scale multimodal tasks (Yue et al., 2025) driving rapid progress. Combined with advances in large language models (Chowdhery et al., 2023; OpenAI, 2023), these systems now achieve near- or superhuman performance on many text-based QA tasks.

However, this success does not extend uniformly to audio question answering (Audio QA), where models must reason over complex and often indirect auditory signals. Despite progress in speech recognition (Baevski et al., 2020), audio tagging (Kong et al., 2020), and multimodal modeling (Guzhov et al., 2022), current systems still struggle to reason about audio beyond surface-level recognition.

Existing benchmarks fail to capture this gap. Datasets such as VGGSound (Chen et al., 2020) focus on closed-set classification, while captioning datasets like Clotho (Drossos et al., 2020) emphasize description rather than reasoning. Many Audio QA datasets further rely on synthetic scenes or templated questions (Fayek and Johnson, 2020; Li et al., 2022), enabling models to exploit language priors or shallow cues rather than true auditory grounding (Section 2).

A key limitation is the absence of systematic human benchmarks, making it difficult to assess whether models are approaching human-level auditory understanding. Recent work has highlighted the need for more reliable evaluation frameworks such as Item Response Theory (Lalor et al., 2019, IRT), which jointly models question difficulty and participant ability.

This work introduces AUDITA (Section 3), a large-scale benchmark of human-authored audio trivia questions sourced from real-world domains including quizzes, media, and cultural knowledge. Unlike prior datasets, AUDITA avoids synthetic audio and templated formats, focusing instead on crafted questions that require multi-cue auditory reasoning and world knowledge. Under controlled human evaluation (Section 4), expert participants have strong accuracy (best-per-category: 69.17% free-form, 86.67% MCQ), while state-of-the-art audio and multimodal models lag far behind, with accuracies under 8.86%. This indicates a substantial gap in auditory reasoning capability.

[1]The codebase and data are available at https://github.com/Pinafore/audio_data, and https://huggingface.co/datasets/TasnimKabir12/audita-audio

Findings of the Association for Computational Linguistics: ACL 2026, pages 25922–25951
July 2-7, 2026 ©2026 Association for Computational Linguistics

We further analyze this gap using IRT (Hambleton et al., 1991), revealing a clear separation between human and model ability distributions. Our analysis also surfaces common issues in existing benchmarks (Section 5): ambiguity, underspecified answers, and shortcut-prone designs, which can inflate model performance without requiring true auditory grounding.

Our contributions are: (1) **AUDITA**, a benchmark for challenging audio QA requiring multi-cue reasoning; (2) **human baselines** demonstrating a large and consistent human–model gap; and (3) **IRT-based analysis** providing fine-grained insights into difficulty, reasoning structure, and model failure modes.

2 Why Audio QA Requires Better Questions

Recent advances in text-based question answering (QA) and large language model training have highlighted that dataset quality remains a primary bottleneck despite advances in model scale. Poorly designed questions introduce ambiguity (Min et al., 2020; Li et al., 2025b), false presuppositions (Yu et al., 2023), unreliable supervision (Li et al., 2024), and exploitable shortcuts (Poliak et al., 2018), resulting in inflated model performance and unstable evaluations. Li et al. (2025b); Shi et al. (2025) show that a substantial fraction of real-world questions remain inherently ambiguous or underspecified, often leading to apparent hallucinations that stem from dataset limitations rather than model failure. Consequently, contemporary QA research has increasingly emphasized careful question design, ambiguity-aware evaluation, and human calibration to ensure benchmarks truly test reasoning rather than dataset artifacts.

Many of these failure modes identified in text QA also appear in Audio QA datasets. We analyze these issues through Item Response Theory (Hambleton et al., 1991, IRT), a classical framework for modeling question difficulty and discrimination, which captures how effectively questions distinguish between high and low-ability models or annotators (Appendix Section A). Rodriguez and Boyd-Graber (2021) and Lalor et al. (2024) have applied IRT to benchmark and leaderboard analysis, providing a principled mechanism for diagnosing dataset quality. This diagnostic perspective aligns with more recent efforts toward adaptive and fluid benchmarking that emphasize maintaining meaningful evaluation as models improve (Hofmann et al., 2025). Appendix Table 8 and 9 summarize common issues across several popular Audio QA datasets, illustrating these problems with concrete examples and IRT-based evidence. Ambiguity, a central challenge extensively studied in (Min et al., 2020), occurs when questions admit multiple plausible answers or interpretations. For example, in Clotho-Audio QA (Drossos et al., 2020), *"What animal is making the sound?"* receives different yet valid responses such as *bird* or *dog*, as both are in the corresponding clip reflecting overlapping audio sources or vague wording. This ambiguity results in low discrimination scores (e.g., VGGSound QA, Appendix Table 8, row 1) and low feasibility.

False presuppositions, well-documented in text QA (Yu et al., 2023), also appear in Audio QA datasets such as FSD50K (Fonseca et al., 2022). These questions assume patterns not grounded in the audio, leading to inconsistent responses; correspondingly, IRT reveals low discrimination, indicating poor ability to differentiate annotator or model competence (Appendix Table 9).

Weak grounding and shortcut learning are also prevalent, especially in caption-derived datasets like AudioCaps QA (Kim et al., 2019) and AudioSet QA (Gemmeke et al., 2017). Questions such as *"Is someone laughing?"* exhibit low discrimination, as they can often be answered without audio, undermining evaluation validity (Appendix Table 8).

Underspecified answer keys (e.g., *dog*, *puppy*, *hound*) in datasets like MUSIC-21 QA (Christodoulou et al., 2025) and Clotho-Audio QA introduce ambiguity, which IRT captures as reduced discrimination and noisy item estimates. Finally, overlapping audio sources (e.g., VGGSound QA, AudioCaps QA) increase ambiguity, producing items with high difficulty but low discrimination— hard yet uninformative questions (Appendix Table 7).

3 Audio Questions for Triva Experts

We introduce a benchmark of human-authored audio questions designed by knowledgeable experts to test auditory and multimodal reasoning skills. This approach aligns with what Rogers et al. (2023) call the "probing" paradigm and what Rodriguez and Boyd-Graber (2021) term the Manchester paradigm, where questions are crafted explicitly to evaluate human-level understanding. While

Dataset	Total Qs	Unique Clips	Avg Q Len	Avg Dur (s)	Dur Range (s)
VGGSound	∼200K	∼200K	12	10	10
Clotho-AQA	11946	1991	14	∼22	15–30
AudioCaps QA	∼50K	∼50K	∼10	∼5	0.5–10
AUDITA	9690	8713	12.47	36.98	0.42–478.33

Table 1: Summary statistics describing the scale and structure of the AUDITA benchmark compared to other Audio QA datasets.

probing-style questions are common in text-only QA datasets such as Kabir et al. (2024) and TriviaQA (Joshi et al., 2017), they remain underexplored in the audio domain.

We scrape questions from publicly available online audio tests, reflecting a broader practice across various communities– such as educators, trivia enthusiasts, and researchers– that release question sets to facilitate human study and benchmarking. This openness aligns with the Manchester paradigm's emphasis on carefully crafted questions designed to probe deep understanding. However, unlike text-based questions, Audio QA requires additional effort to segment continuous audio content into discrete, independently answerable questions that can be posed one at a time to humans or machines. For instance, many examples have text instructions that apply to all of the following examples: e.g., *"how many wheels do each of the following vehicles have"*, followed by the sound that each of the vehicles makes.

In total, AUDITA contains 9690 audio-question pairs, providing a valuable resource for studying audio question answering. Existing audio QA datasets differ substantially in scale, audio structure, and annotation philosophy, spanning large weakly-purposed corpora like VGGSound, human-curated benchmarks such as Clotho-AQA, and caption-derived QA sets like AudioCaps QA (Table 1). Despite their widespread use, these datasets are not systematically standardized: key properties such as question formulation, audio duration regimes, and annotation consistency vary significantly and are often incompletely reported. This lack of controlled design leads to confounded comparisons across methods and limits meaningful evaluation of auditory reasoning, motivating more rigorously constructed and diagnostically transparent benchmarks such as AUDITA.

3.1 AUDITA Data Sources

We describe the sources used in AUDITA, which span curated audio collections and structured trivia-style datasets.

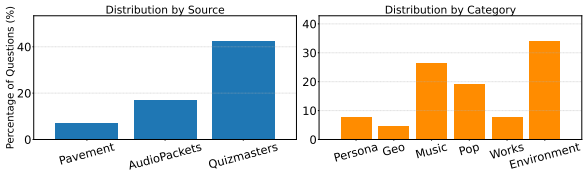

Figure 1: Distribution of questions across sources and categories in the dataset. The dataset exhibits diverse and rich coverage across both sources and categories, ensuring broad domain representation and supporting robust evaluation across varied audio question answering settings.

Quizmasters Website The Quizmasters website (Quizmasters, 2025) hosts curated collections of short audio clips organized by auditory skill category, originally created for pub quiz-style challenges, but without accompanying questions. These collections are designed to probe specific listening abilities, such as recognizing transformed audio (e.g., reversed or filtered signals) or identifying less common musical material. We convert these collections into an audio question answering format by attaching human-authored questions aligned with each category's intended challenge. In addition to question–answer pairs, we retain clip-level metadata such as duration and sampling rate.

Audio Pyramidal Trivia and PAVEMENT Pyramidal question structure is established best practice for human question answering (HQA) (Boyd-Graber and Börschinger, 2020), as it measures system knowledge by revealing clues incrementally from obscure to obvious. This principle extends naturally to audio: a pyramidal audio question might begin with an isolated instrument passage, progress through a fuller excerpt, and conclude with a recognizable theme– rewarding systems that identify the answer from minimal acoustic context rather than only after highly identifying cues. Audio Pyramidal Trivia and PAVEMENT (Rodriguez et al., 2019; Pavao, 2025) both adopt this paradigm, structuring questions as sequences of progressively more informative audio clips. While earlier collections focus primarily on music (with some extensions to film and television), PAVEMENT (Pavao, 2025) adds human-written competitive questions and richer annotations including acceptable answers and clarifications. Conventional QA benchmarks, by contrast, present questions as atomic, non-incremental units, leaving this evaluative dimension unaddressed. By retaining PAVEMENT's pyramidal structure and segmenting each question into individual audio

question pairs, our evaluation captures a difficulty axis that flat CQA formats cannot.

To ensure correctness and consistency, we perform dataset cleaning and normalization, including verifying audio–question alignment and standardizing formatting. Figure 1 and Appendix Table 5 present the distribution of questions by source.

3.2 Data Preparation

AUDITA is constructed through a three-stage pipeline: alignment, normalization, and categorization to produce consistent and human-readable audio-question-answer triples for evaluation. The process ensures that questions are correctly paired with their corresponding audio clips, removes formatting artifacts, and standardizes answer representations to reduce variability across sources. Details of the scraping, cleaning, and normalization procedures are provided in Appendix B.

Categorization and consolidation. Because our dataset and audio QA benchmarks more broadly span diverse auditory phenomena and reasoning skills, we organize questions into semantic categories. This structuring allows participants to answer queries relevant to their expertise, making the task more approachable and efficient for skilled annotators. It also enables fine-grained analysis of model behavior across domains (e.g., music recognition versus environmental sound reasoning).

We assign each question to a hierarchical taxonomy of six high-level categories and 26 subcategories using GPT-4o-mini. For usability in the human-centered QA task, we collapse this hierarchy into six interpretable categories exposed to participants: *Cultural Geography in Sound*, *Name The Music: Songs, Artists & Composers*, *Who's Who? Name That Persona*, *Elements of Musical Works*, *Pop Culture and Media*, and *Environmental and Acoustic Sound Recognition*. This structure is applied uniformly across all data sources (Figure 1 and Appendix Table 6).

MCQ generation We create MCQ variants with one correct answer and three AI-generated distractors validated by human annotators. The distractors are designed as plausible alternatives, sharing surface-level cues with the correct answer while differing in the key auditory evidence required for identification. MCQ variants help reveal both accuracy patterns and question quality, particularly in settings where models are not performing well and free-form responses may be influenced by language priors or generation artifacts. By requiring selection among closely related options, MCQs encourage discrimination rather than free-form generation or binary decisions, making them useful for diagnosing whether errors arise from model limitations or from the discriminative strength of the question.

For each question, we first determine the semantic type of the correct answer (e.g., musical artist, actor/actress, etc.), and then create three distractors that matched key attributes of the gold answer to ensure plausibility. For instance: If the correct answer was a **musical artist**, distractors match gender, genre, and era; be real existing artists; not be the same individual or an alias; and not violate entity type constraints (e.g., no band if the correct answer is a solo artist). If the correct answer was an **actor/actress**, distractors match gender, era, profession, and accent; be real individuals; and not be aliases of the correct answer.

Each row was generated independently to avoid cross-item leakage or systematic reuse patterns. After initial creation, a second author independently reviewed all distractors to verify that constraints were satisfied; all entities were valid and distinct; and no option was trivially eliminable.

3.3 Dataset Composition

AUDITA is constructed from two sources: (i) original human-authored audio clips with associated questions, and (ii) external benchmark datasets. In total, it contains 9690 questions, including 6460 human-authored questions and 3230 questions from external benchmarks. The human-authored portion is built on curated audio clips sourced from *Pavements*, *Audio-Packets*, and *Quizmasters*. This includes 2322 pyramidal-style questions (673 from *Pavements* and 1649 from *Audio-Packets*) and 4138 trivia-style questions from *Quizmasters*. All human-authored questions are closed-ended with discrete, verifiable answers.

The external portion comprises 2907 questions from OpenAudio QA (90%) and 323 questions from ClothoAudio QA (10%) (Gong et al., 2023b; Lipping et al., 2022). The external set contains 1205 closed-ended questions (882 from OpenAudio QA and 323 from ClothoAQA) and 2025 open-ended questions (all from OpenAQA) that require semantic evaluation rather than exact string matching. For OpenAQA, we apply the original filtering procedure from Gong et al. (2024), re-

moving 18.65% of instances where the generation pipeline itself flags the question as unanswerable (e.g., responses containing "cannot be determined" or "unclear"), leaving only questions with committed answers (details in Appendix Section C).

We include external datasets not as core evaluation targets but as *reference points* to contextualize the behavior of our human-authored data. While OpenAQA and ClothoAQA are widely used audio QA benchmarks, they lack human validation, making it difficult to directly assess how close models are to human-level performance; including human responses enables this direct comparison. These datasets also provide an important contrast in difficulty and structure: external questions yield higher human accuracy and smaller human-model gaps, reflecting reliance on short, perceptual, or caption-derived cues. In contrast, our human-authored questions are explicitly designed to be more challenging and reasoning-intensive.

External datasets play a complementary role rather than serving as the primary benchmark. All analyses are disaggregated by source, and our main claims about human–model gaps focus on the human-authored subsets. ClothoAQA, despite structural limitations noted in Section 2, is included as an independently crowd-sourced dataset that provides a complementary human-annotated reference point beyond OpenAQA. Overall, the external portion serves to *anchor* our evaluation– enabling comparability with prior work while highlighting limitations of existing benchmarks relative to the reasoning demands of AUDITA.

Across the dataset, in the human evaluation described in the next section, we collect 1517 human guesses, individual answer attempts by participants, providing a reliable set of judgments to benchmark model performance. Our evaluation framework leverages these human guesses to jointly model question properties and participant abilities.

4 How hard is AUDITA for Humans and Computers

We evaluate the proposed dataset using both state-of-the-art audio-language models and human participants to characterize its difficulty and suitability for auditory reasoning. We collect both free-form and multiple-choice (MCQ) responses under identical input conditions, enabling complementary assessment of generative and decision-based reasoning.

4.1 Human Evaluation

To ground performance in human capability and better understand the intrinsic difficulty of AUDITA, we recruit participants from online quiz and trivia communities, who regularly engage with general-knowledge and audio-based trivia content. This provides a population with strong familiarity with trivia-style reasoning while remaining non-expert in the specific benchmark content.

We ask participants to evaluate the same audio clips and questions under identical conditions, without access to transcripts or external resources, ensuring that responses reflect purely auditory understanding. We elicit responses in both multiple-choice and free-response settings. Full instruction are provided in Appendix Section F.

Answer Evaluation Free-form answers are evaluated using the PEDANTS (Li et al., 2024) framework for semantic equivalence, allowing for variation in phrasing while preserving correctness. In contrast, MCQ responses are evaluated through selection among predefined answer options, providing a controlled setting for discriminative reasoning (Appendix Section D).

4.2 Models

We evaluate 18 models, including both open-source audio-multimodal systems and proprietary large models (GPT-4o and Gemini 2.5 Pro) to provide coverage across capability scales. The open-source models consist of 16 systems with mid-scale language backbones (approximately 4B–13B parameters), grouped into three categories. **Omnimodal models** (6) support unified understanding across text, audio, vision, and video with both text and speech generation. **Audio–language models** (4) focus on audio understanding with text-only outputs, while **speech-capable models** (6) emphasize speech recognition and generation, including both speech-first and modular architectures. GPT-4o and Gemini 2.5 Pro are included as representative state-of-the-art proprietary systems for comparison against recent large-scale multimodal models.

This taxonomy reflects the diversity of AUDITA questions: some require speech or lyric understanding, others test purely acoustic reasoning over music or environmental sounds, and some benefit from broader multimodal context. Evaluating across capability groups enables analysis of whether failures are systematic or capability-specific, analogous to how humans identify songs via lyrics or melody.

System	Free-from Accuracy	θ	SD	95% CI	MCQ Accuracy	θ	SD	95% CI
Humans	32.13	0.05	0.26	[-0.001, 0.101]	60.16	0.08	0.25	[0.055, 0.105]
Models (avg)	8.86	-2.91	0.55	[-3.097, -2.723]	15.65	-2.45	0.54	[-2.636, -2.264]

Table 2: Human vs model aggregate performance with IRT-based ability estimates and uncertainty intervals. Humans substantially outperform models across both text and MCQ settings, with a large and consistent gap in estimated ability (θ), indicating a strong human–model performance disparity beyond raw accuracy.

All models are evaluated using publicly released checkpoints (for open-source systems) and API-based inference (for proprietary models), with recommended settings and no task-specific fine-tuning. Full model details are provided in Appendix E.

4.3 Input Representation

To fairly evaluate both humans and computational models on our audio question answering task, we ensure consistent presentation of audio and questions. Each example consists of an audio clip paired with a natural-language question, presented verbatim without templating or normalization. For human participants, audio clips are provided in standard formats (.mp3, .wav) to ensure consistent playback across devices.

For computational models, audio inputs are uniformly preprocessed: audio is converted to mono and resampled to the sample rate expected by each model. Depending on the model's input requirements, we supply either raw waveform audio or alternative audio representations such as log-mel spectrograms or codec tokens, following each model's prescribed preprocessing pipeline.

5 Human–Model Performance Gaps with Accuracy and IRT

Humans have a substantial and consistent lead over state-of-the-art audio-language models on AUDITA across both free-form and multiple-choice settings (human: 32.13% vs. models: 8.86% in free-form answers; human: 60.16% vs. models: 15.65% in MCQ). However, accuracy alone obscures important structure in QA benchmarks.

While these accuracy gaps clearly indicate strong human advantages, they do not capture the quality or diagnostic value of individual questions in AUDITA. In particular, questions vary substantially in how effectively they distinguish between high- and low-ability respondents.

We use **Item Response Theory (IRT)** as a principled framework for analyzing human and model performance and diagnosing evaluation quality (Baker and Kim, 2004; Embretson and Reise, 2013). Widely used in standardized testing and leaderboard construction, IRT enables comparison across heterogeneous populations by normalizing for question difficulty (Rodriguez and Boyd-Graber, 2021). This is particularly important in AUDITA, where items vary substantially in acoustic complexity and reasoning demands. By jointly modeling difficulty and discrimination, IRT supports analysis beyond aggregate accuracy and identifies which questions are most informative. For example, in AUDITA high-discrimination items include entity identification from audio clips, while low-discrimination items often involve underspecified prompts such as language identification or continuation-based queries (Appendix Table 14).

Model	Free-from Acc.	θ	Rank	MCQ Acc.	θ	Rank
Gemini 2.5 Pro (Comanici et al., 2025)	17.02	-1.66	1	39.18	-1.40	1
GPT-4o (Hurst et al., 2024)	14.87	-1.97	2	34.89	-1.62	2
Qwen2.5-Omni (Xu et al., 2025a)	10.01	-2.16	3	21.02	-1.76	4
AudioGPT (Huang et al., 2024)	8.98	-2.31	4	23.49	-1.61	3
OpenOmni (Luo et al., 2025)	7.99	-2.45	5	16.01	-2.06	6
Audio-Flamingo (Kong et al., 2024)	7.77	-2.49	6	13.99	-2.11	8
Phi-4-Multimodal (Abouelenin et al., 2025)	7.49	-2.54	7	15.73	-2.09	7
Qwen3-Omni (Xu et al., 2025b)	6.87	-2.62	8	18.89	-2.04	5
LTU-AS (Gong et al., 2023a)	6.69	-2.62	9	13.81	-2.11	9
Qwen-2 Audio (Chu et al., 2024)	6.53	-2.70	10	13.72	-2.31	10
Baichuan-Omni-1.5 (Li et al., 2025a)	6.49	-2.71	11	13.63	-2.43	11
VITA-1.5 (Fu et al., 2025)	5.76	-2.81	12	12.59	-2.71	12
Mini-Omni2 (Xie and Wu, 2024)	4.85	-3.03	13	8.74	-2.78	14
SpeechGPT (Zhang et al., 2023)	3.99	-3.28	14	12.38	-2.71	13
SALMONN-2 (Tang et al., 2025)	2.77	-3.60	15	5.68	-3.31	17
SALMONN-2+ (Tang et al., 2025)	2.62	-3.61	16	6.28	-2.89	16
MU-LLaMA (Liu et al., 2024)	2.19	-3.61	17	7.48	-2.81	15
SALMONN (Tang et al., 2024)	1.99	-4.00	18	4.18	-3.40	18

Table 3: Full model comparison on text and MCQ settings using accuracy and IRT-based ability estimates. State-of-the-art proprietary models (e.g., Gemini 2.5 Pro and GPT-4o) lead across both settings, while open-source models exhibit a large performance spread, with consistent rankings across accuracy and IRT-based ability estimates.

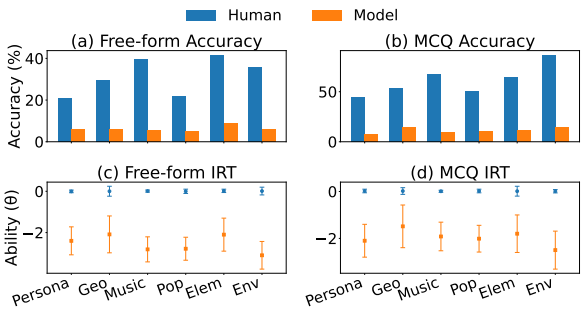

Figure 2: Category-level accuracy and average item difficulty for humans and models on free-form and multiple-choice questions. Bars show accuracy; points show mean IRT difficulty ± one standard deviation. Higher difficulty correlates with lower accuracy for both, but models' performance declines more sharply, revealing systematic category-specific gaps.

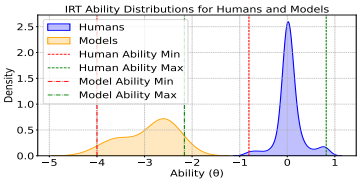

Figure 3: Distributions of IRT ability (θ) for humans (blue) and models (orange) are shown on a shared scale. Kernel density estimates highlight that humans cluster at higher θ, with dashed lines indicating each group's range, revealing a clear latent ability gap despite model variability.

Table 2 and Table 3 report accuracy alongside IRT ability estimates for humans and models. In addition to measuring performance, IRT highlights poorly diagnostic items: low-discrimination questions fail to separate high- and low-ability respondents, while extremely easy or difficult items provide limited signal for comparison. Humans substantially outperform all models under both metrics, and system rankings remain broadly consistent across accuracy and ability.

Both humans and models perform well on easier questions, while the performance gap widens with increasing difficulty, with humans remaining more robust on harder items. Easier questions typically contain clear audio cues, familiar sounds, or require limited temporal integration, corresponding to lower IRT difficulty (b) values. Earlier benchmarks tend to exhibit a narrower difficulty range, where models achieve higher accuracy and show less variation in estimated ability (θ). In contrast, AUDITA spans a broader difficulty spectrum, producing larger and more consistent human–model gaps and better exposing model limitations. Additional item- and dataset-level breakdowns are provided in Appendix Figure 4 and Table 10.

To localize this gap, Figure 2 presents category-level accuracy alongside mean item difficulty. Categories involving music and environmental sound consistently show strong human performance but weak model performance. For example, in Elements of Musical Works, humans reach 41.6% (free-form) and 64.8% (MCQ), while models remain low at 8.7% and 11.3% (IRT ability estimate θ:0.03 vs. -2.1). Similar gaps appear in Name The Music (39.4% / 67.6% vs. 5.4% / 9.7%) and Environmental Sound Recognition (86.7% vs. 14.9% MCQ). Overall, consistent gaps across categories indicate that failures stem from intrinsic auditory reasoning difficulty rather than annotation or evaluation noise. Figure 3 shows the IRT ability distributions for humans and models on a shared scale. Models concentrate in a narrower, substantially lower ability range than humans, indicating reduced robustness as item difficulty increases.

Setting	Input Modalities	Accuracy
Text-only	Question only (no audio)	0.0001%
Transcript-only	Transcript (no audio)	4.26%
Multimodal	Audio + text	8.86%
Text-only MCQ	Question + options (no audio)	1.29%
Transcript-only MCQ	Transcript + options (no audio)	7.05%
Multimodal MCQ	Audio + text + options	15.65%

Table 4: Near-zero and lower performance in text-only and transcript-only settings and gains from multimodal input confirm that AUDITA cannot be solved using textual priors alone and requires audio understanding.

Category-level analysis shows that environmental sounds and complex music remain particularly challenging for models, while humans remain strong across these domains. Models often confuse acoustically similar clips or fail to capture long-range temporal cues, performing relatively better in MCQ settings but still showing substantial gaps. Overall, IRT-based analysis confirms that models struggle most on high-difficulty items, highlighting limitations in robust auditory reasoning.

Baseline Analysis and MCQ Performance To isolate the contribution of different modalities in AUDITA, we evaluate text-only, transcript-only, and multimodal settings under both free-form and MCQ formats (Table 4). Text-only settings achieve near-zero accuracy, confirming that the benchmark cannot be solved using linguistic priors alone. Even with transcripts, performance remains low (4.26%), as many examples contain minimal or no spoken content (e.g., music, environmental sounds, speaker identity). In contrast, multimodal inputs improve performance relative to text-only variants, demonstrating that non-zero accuracy depends on access to acoustic information.

In the MCQ setting, accuracy (15.65%) falls below the 25% chance level. This does not indicate random guessing; instead, it reflects systematic miscalibration, where models consistently prefer plausible but incorrect options due to fine-grained acoustic distinctions between distractors (Zhao et al., 2021; Wang et al., 2025c). Similar behavior has been observed in prior multimodal evaluation settings, where models rely on spurious correlations or superficial cues that misalign with the underlying evidence (Kavumba et al., 2022; Wu et al., 2025). These results show that AUDITA evaluates genuinely audio-grounded understanding rather than text-based reasoning.

Interpreting human accuracy AUDITA consists of open-ended audio trivia with a large effective answer space, making chance performance in free response effectively zero. In multiple-choice, chance accuracy is $1/K$ (e.g., 25% for K=4), yet humans substantially exceed this baseline under identical conditions. Free-form-answer scoring is intentionally strict; many items admit plausible near-misses (e.g., confusing franchise installments or covers vs. originals)–so even moderate free-form response accuracy reflects meaningful auditory reasoning rather than guessing, consistent with human baselines reported by MMAU (Sakshi et al., 2024).

5.1 Factual Knowledge vs. Audio Understanding

Some questions in AUDITA-for example, *"Name the artist who recorded this song"* or *"What film is this theme from?"*- require both perceptual recognition and world knowledge. To disentangle these factors, we compare two model settings already available in our experiments: *Question + Transcript* (Whisper output) and *Question + Raw Audio* (see Appendix J.1 for full details and examples).

This comparison allows us to interpret performance differences diagnostically. Consider *"Name the songwriter"* with the answer *Joni Mitchell(River)*. Here, the transcript contains lines from the song, enabling correct identification, whereas the raw audio alone does not provide enough cues, representing a perceptual limitation. In contrast, *"You are listening to the voice of a fictional character; what is this character's name?"* with the answer *Frank Spencer*, both transcript and audio fail, reflecting a knowledge limitation. Finally, in cases such as *"Name the composer"* with the answer *Carl Orff*, the distinctive choruses and instrumental timbres in the audio allow the model to succeed, while the transcript contains no usable information; this highlights reliance on non-verbal acoustic cues.

Across all items, accuracy is substantially higher in the raw-audio condition (8.86%) than in the transcript condition (4.26%). If performance were primarily driven by textual or factual recall, transcript-based accuracy would match or exceed audio-based results. This gap indicates that transcripts often omit or distort critical acoustic information, making non-verbal audio cues essential for identification. This difference shows that failures cannot be explained solely by missing world knowledge, but reflect genuine limitations in audio understanding.

5.2 Failure Case Analysis

We analyze model failures in AUDITA through three error modes: knowledge-based, perceptual, and audio-cue reasoning failures. Knowledge-based errors dominate at 78.23%, while the remaining 21.77% involve failures where auditory perception or acoustic cues play a decisive role. A key finding is that even when models have access to audio signals, they frequently fail on tasks requiring fine-grained auditory reasoning combined with cultural or entity-level knowledge. For example, music recognition and cultural media questions often require identifying composers, artists, or works from subtle acoustic signatures such as melody structure, timbre, orchestration, or vocal style. In *Name The Music*, humans achieve 39.4% (free-form) and 67.6% (MCQ) accuracy, while models remain at 5.4% and 9.7%, highlighting a large gap in auditory grounding and entity recognition. Similarly, in *Environmental and Acoustic Sound Recognition*, humans reach 35.7% MCQ accuracy compared to 14.9% for models, indicating that humans can reliably infer sound sources from acoustic structure, whereas models struggle to form stable perceptual representations of environmental sounds.

Perceptual and audio-cue errors also highlight limitations in how models exploit acoustic information. Some failures stem from weak use of discriminative audio structure, while others reflect difficulty capturing temporal patterns, texture, and non-speech cues absent from text. These cases show that success requires not only linguistic reasoning but also direct acoustic grounding and temporal pattern recognition in time-varying signals.

Knowledge-based failures remain the largest category, but they often co-occur with weak or underspecified acoustic evidence. Many questions require identifying culturally grounded entities such as operas, songs, or fictional characters, where success depends on linking auditory cues with prior world knowledge. When either component is weak, performance degrades sharply. These results suggest that errors in AUDITA are not solely due to missing factual knowledge. Instead, they stem from failures to jointly integrate auditory perception (e.g., timbre, rhythm, and sound structure) with entity-level reasoning over cultural and factual knowledge. This interaction explains why models fail even with access to audio, highlighting a fundamental gap in auditory grounding compared to human listeners (details in Appendix Section J.2).

6 Related Work

We review prior work in audio QA, multimodal and adversarial evaluation, and identify key limitations motivating AUDITA.

Audio QA datasets Audio Question Answering has evolved from synthetic benchmarks to more natural but still limited human-sourced datasets. Early works such as CLEAR (Lin et al., 2021) generate synthetic acoustic scenes with compositional questions, enabling controlled evaluation but limiting realism. DAQA (Fayek and Johnson, 2020) similarly uses constrained sound compositions with yes/no or count-based questions, reducing reasoning to structured classification. More natural datasets such as ClothoAQA (Drossos et al., 2020) and Music-AVQA (Li et al., 2022) introduce crowd-sourced questions over real audio or video. However, they still exhibit strong linguistic priors and template bias, allowing models to succeed with limited audio grounding and often overestimating performance (Appendix Table 17). The DCASE 2025 Audio QA challenge (Yang et al., 2025) expands domain coverage but remains largely multiple-choice and non-adversarial. Prior work further shows sensitivity to textual artifacts and shortcut cues in audio-language models (Wang et al., 2025a), suggesting current benchmarks do not reliably measure robust auditory reasoning.

Adversarial and multimodal QA Outside audio, adversarial QA exposes model weaknesses via human-in-the-loop construction of challenging examples. In vision-language QA, methods such as Adversarial VQA and Dynabench (Sheng et al., 2021; Kiela et al., 2021) iteratively generate harder questions via model feedback. Recent analyses show that many benchmarks fail to induce true adversarial behavior despite appearing challenging (Sung et al., 2025). In text QA, adversarial datasets such as (Wallace et al., 2019) reduce shortcut reliance and improve robustness evaluation. Multimodal QA benchmarks including VQA (Antol et al., 2015), CLEVR (Johnson et al., 2017a), GQA (Hudson and Manning, 2019), and MultimodalQA (Talmor et al., 2021) address bias and compositionality via balancing and structured grounding. Video and audio-visual QA datasets like AVQA (Yang et al., 2022) and Music-AVQA (Li et al., 2022) extend these ideas temporally but still rely largely on templated or crowd-sourced questions (Appendix Section H).

Positioning relative to prior Audio QA benchmarks Prior Audio QA benchmarks focus on recognition-style tasks: closed-set event labeling (e.g., *dog barking*, *siren*), caption- or metadata-derived QA (e.g., *"What instrument is playing?"*), synthetic or templated mixtures, and speech-centric pipelines reducing audio QA to ASR followed by text QA. As summarized by MMAU (Sakshi et al., 2024), these settings test information extraction with limited reasoning. In contrast, AUDITA targets a regime requiring of grounding audio to real-world entities and integrating long-range cues. Instead of identifying sounds, it asks questions such as *"Which song is this clip from?"*, *"Which film does this score belong to?"*, or *"Which artist is associated with this segment?"*. Answering these requires audio-to-referent grounding and reasoning over distributed acoustic cues (e.g., melody, timbre, context), shifting the task from surface recognition to entity-level inference.

Item Response Theory in NLP evaluation Item Response Theory (IRT), originally developed in psychometrics, has been increasingly adopted in NLP for evaluating datasets and models. Lalor et al. (2016); Rodriguez and Boyd-Graber (2021) show that modeling item difficulty and discrimination provides a more informative alternative to aggregate accuracy and enables comparison across heterogeneous examples and systems. IRT has also been used to scale evaluation using model-generated responses and to diagnose dataset quality by identifying uninformative or low-discrimination items (Lalor, 2020).

7 Conclusion

We present AUDITA, a benchmark of human-authored audio questions grounded in real-world settings, designed to evaluate genuine auditory reasoning. Our results show a clear and persistent gap between human and model performance, with IRT analysis revealing limitations in handling acoustic cues, temporal structure, and context. These findings suggest that scaling alone is insufficient.

Future work should focus on improving audio-language alignment, strengthening entity-level reasoning, and enabling multi-step inference over complex audio inputs, potentially through retrieval-augmented approaches that incorporate external knowledge sources. More fine-grained diagnostic frameworks can further help isolate and address these failures.

8 Limitations

Our evaluation targets out-of-the-box performance of locally runnable, open-checkpoint models in a mid-scale regime, and we do not claim to cover the full space of proprietary or cloud-only systems. Larger closed models may improve absolute accuracy, but the scale question is discussed directly in Appendix I.1, including why scale alone is unlikely to close the gap we observe on a difficult, human-authored benchmark. We also do not run full scaling sweeps across cloud-scale systems because evaluating 9,690 items with audio inputs would impose substantial cost.

We intentionally evaluate end-to-end model behavior without external tool augmentation. In particular, we do not benchmark pipelines that add ASR plus retrieval, music fingerprinting, database lookups, or web search. These systems are relevant in practice and may reduce errors on referent-linking items, but they measure a different capability than the audio-grounded reasoning we aim to isolate. Relatedly, we evaluate all systems in an audio+question to text-answer setting for consistency, even when a model can generate speech, which may understate the value of speech-first interaction designs in real deployments.

Our psychometric analysis depends on the breadth and quality of human responses. While IRT helps separate item difficulty from responder ability, it can still be affected by annotator variability and by items where the audio is noisy, overlapping, or underspecified. Human participants are also not uniformly "trivia experts," so aggregate human accuracy should be interpreted as a baseline rather than a ceiling. Finally, although AUDITA is deliberately sourced from real-world domains, it is still shaped by the distribution of publicly available audio trivia and by English-centric question writing, which may limit generalization to other languages, accents, and niche audio domains.

AUDITA contains short audio clips (averaging 37 seconds) from publicly available trivia sources, following practices established by multimedia QA benchmarks such as TVQA (Lei et al., 2018). We will release stable metadata and acquisition scripts to support reproducibility.

9 Ethical Considerations

Human evaluation was conducted under Institutional Review Board (IRB) approval and informed consent. We collected participant email addresses solely to deliver remuneration. Emails are stored securely, are not used for analysis, and are not linked to response data in the released benchmark or reported results. Aside from compensation logistics, we do not collect additional personally identifying information, and we analyze results in anonymized form. The audio content used in AUDITA is sourced from publicly available datasets or permissively licensed materials, and is used for research purposes. We do not redistribute restricted content beyond derived annotations required for benchmarking.

Acknowledgement

This work is supported by TRAILS (Institute for Trustworthy AI in Law & Society) (CNS-2150382). We thank the University of Maryland Institute for Advanced Computer Studies (UMIACS) for continuous support and computational resources. We also thank the reviewers for their valuable comments and suggestions, which helped improve the clarity and quality of this work. We are grateful to all participants who contributed to the human evaluation. The following participants explicitly consented to being acknowledged by name: Daniel Kim, Drew Scheeler, Eve Nuria Fleisig, Forrest Weintraub, Hemanth Nandakumar, Jason Christopher, Mohammed Afaan Mohammed Arif Ansari, Nathan Zhao, Nishant Balepur, Raymond Kimball, Sara DelVillano, and Stefany Meyer. Their annotations and responses were essential for establishing reliable human performance benchmarks and validating the quality and difficulty of the dataset. We especially appreciate the time and effort they invested in engaging with challenging audio-based questions. We also acknowledge the authors of prior datasets and resources used in this work for making their data and tools publicly available.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, and Dongdong Chen and. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *CoRR*, abs/2503.01743.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt*

Lake City, UT, USA, June 18-22, 2018, pages 4971–4980. Computer Vision Foundation / IEEE Computer Society.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Frank B Baker and Seock-Ho Kim. 2004. *Item response theory: Parameter estimation techniques*. CRC press.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 721–725. IEEE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Anna-Maria Christodoulou, Kyrre Glette, Olivier Lartillot, and Alexander Refsum Jensenius. 2025. Musiqal: A dataset for music question-answering through audio-video fusion. *Trans. Int. Soc. Music. Inf. Retr.*, 8(1).

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and Evan Rosen. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261.

Santiago Cuervo and Ricard Marxer. 2024. Scaling properties of speech language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 351–361. Association for Computational Linguistics.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 736–740. IEEE.

Susan E Embretson and Steven P Reise. 2013. *Item response theory for psychologists*. Psychology Press.

Haytham M. Fayek and Justin Johnson. 2020. Temporal reasoning via audio question answering. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2283–2294.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: an open dataset of human-labeled sound events. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:829–852.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yifan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. VITA-1.5: towards gpt-4o level real-time vision and speech interaction. *CoRR*, abs/2501.01957.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James R. Glass. 2023a. Joint audio and speech understanding. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in

visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 976–980. IEEE.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory*, volume 2. Sage.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2025. Fluid language model benchmarking. In *Second Conference on Language Modeling*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 23802–23804. AAAI Press.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and Aleksander Mkadry. 2024. Gpt-4o system card.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3008–3017. IEEE Computer Society.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Tasnim Kabir, Yoo Yeon Sung, Saptarashmi Bandyopadhyay, Hao Zou, Abhranil Chandra, and Jordan Lee Boyd-Graber. 2024. You make me feel like a natural question: Training QA systems on transformed trivia questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20486–20510, Miami, Florida, USA. Association for Computational Linguistics.

Pride Kavumba, Ryo Takahashi, and Yusuke Oda. 2022. Are prompt-based models clueless? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2333–2352.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 119–132. Association for Computational Linguistics.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2880–2894.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew

Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

John P Lalor. 2020. Learning latent characteristics of data and models using item response theory.

John P Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 9–13.

John P Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4248–4258. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: localized, compositional video question answering. *CoRR*, abs/1809.01696.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19086–19096. IEEE.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021. Adversarial VQA: A new benchmark for evaluating the robustness of VQA models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2022–2031. IEEE.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, and Lingfeng Ming and. 2025a. Baichuan-omni-1.5 technical report. *CoRR*, abs/2501.15368.

Zongxi Li, Yang Li, Haoran Xie, and S Joe Qin. 2025b. Condambigqa: A benchmark and dataset for conditional ambiguous question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2288.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan L. Boyd-Graber. 2024. PEDANTS: cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 9373–9398. Association for Computational Linguistics.

Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. 2021. The CLEAR benchmark: Continual learning on real-world imagery. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. *arXiv preprint arXiv:2204.09634*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 286–290. IEEE.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Yangyi Chen, Hamid Alinejad-Rokny, and Fei Huang. 2025. Openomni: Large language models pivot zero-shot omnimodal alignment across language with real-time self-aware emotional speech synthesis. *CoRR*, abs/2501.04561.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5783–5797. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Victor Pavao. 2025. Pavement-Music-Tournaments. *Github*. Online; Available at: `https://github.com/vcpavao/Pavement-Music-Tournaments/`.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the seventh joint conference on lexical and computational semantics*, pages 180–191.

Quizmasters. 2025. The Quizmasters Audio Database. Online; Available at: `https://www.quizmasters.biz/DB/Audio/AMenu.html`.

Pedro Rodriguez and Jordan Boyd-Graber. 2021. Evaluation paradigms in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9630–9642, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pedro Rodriguez and Jordan L. Boyd-Graber. 2021. Evaluation paradigms in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9630–9642. Association for Computational Linguistics.

Pedro Rodriguez and Jordan Lee Boyd-Graber. 2021. Evaluation paradigms in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9630–9642.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan L. Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *CoRR*, abs/1904.04792.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10):197:1–197:45.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *Preprint*, arXiv:2410.19168.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20346–20359.

Zhengyan Shi, Giuseppe Castellucci, Simone Filice, Saar Kuzi, Elad Kravi, Eugene Agichtein, Oleg Rokhlenko, and Shervin Malmasi. 2025. Ambiguity detection and uncertainty calibration for question answering with large language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 41–55, Albuquerque, New Mexico. Association for Computational Linguistics.

Yoo Yeon Sung, Maharshi Gor, Eve Fleisig, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Is your benchmark truly adversarial? AdvScore: Evaluating human-grounded adversarialness. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 623–642, Albuquerque, New Mexico. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-salmonn 2: Captioning-enhanced audio-visual large language models. *CoRR*, abs/2506.15220.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. 2025a. When audio and text disagree: Revealing text bias in large audio-language models. *CoRR*, abs/2508.15407.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. 2025b. When audio and text disagree: Revealing text bias in large audio-language models. *Preprint*, arXiv:2508.15407.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025c. Llms may perform mcqa by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209.

Chengfei Wu, Ronald Seoh, Bingxuan Li, Liqiang Zhang, Fengrong Han, and Dan Goldwasser. 2025. Magic: Evaluating multimodal cognition toward grounded visual reasoning. In *First Workshop on Foundations of Reasoning in Language Models*.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *CoRR*, abs/2410.11190.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin

Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. 2025b. Qwen3-omni technical report. *CoRR*, abs/2509.17765.

Ziqi Xu, Sevvandi Kandanaarachchi, Cheng Soon Ong, and Eirini Ntoutsi. 2025c. Fairness evaluation with item response theory. In *Proceedings of the ACM on Web Conference 2025*, pages 2276–2288.

Chao-Han Huck Yang, Sreyan Ghosh, Qing Wang, Jaeyeon Kim, Hengyi Hong, Sonal Kumar, Guirui Zhong, Zhifeng Kong, Sakshi Singh, Vaibhavi Lokegaonkar, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, Gunhee Kim, Jun Du, Rafael Valle, and Bryan Catanzaro. 2025. Multi-domain audio question answering toward acoustic content reasoning in the DCASE 2025 challenge. *CoRR*, abs/2505.07365.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. AVQA: A dataset for audio-visual question answering on videos. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3480–3491. ACM.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10457–10480. Association for Computational Linguistics.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15757–15773. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. Pmlr.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, et al. 2026. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35085–35093.

This appendix provides detailed documentation of the data sources, processing pipeline, evaluation settings, and extended analyses used to construct and study AUDITA. Section B describes data collection and processing, including source-specific details for Quizmasters and PAVEMENT, as well as extraction, alignment, and normalization procedures. It also includes dataset statistics (Tables 5 and 6) and qualitative analyses of existing audio QA datasets (Tables 7, 8, and 9).

Section D outlines the evaluation settings, including free-response scoring using PEDANTS and multiple-choice evaluation. Section E details the evaluated models, their architectures, training paradigms, and capability groupings. Section H situates AUDITA within prior work on audio QA, multimodal QA, and adversarial evaluation.

Section I.1 examines the role of model scale and validates that our conclusions are not artifacts of model size alone. The main results and diagnostic analyses are presented in Section G, including MCQ behavior, human agreement, task difficulty, and IRT-based evaluation (Section A). Section J.1 and Section J.2 further analyze model failures across modalities, distinguishing perceptual, knowledge-based, and audio-driven errors.

Finally, Section F documents the human evaluation protocol and participant setup, with additional details in Sections F and F.1. Together, these sections provide a comprehensive view of dataset construction, evaluation methodology, and the underlying factors driving model performance on AUDITA.

A IRT Model Specification

Item Response Theory (IRT) is an increasingly common method for discovering gaps between human and machine ability and for identifying problematic examples (Baker and Kim, 2004; Embretson and Reise, 2013).

We adopt the standard **two-parameter logistic (2PL) IRT model**. In this framework, each model (or human group) is treated as a "respondent" with ability parameter θ, and each question is characterized by two parameters:

1. **Difficulty** b
2. **Discrimination** a

The probability that a respondent with ability θ correctly answers item i is given by

$$P(\text{correct} \mid \theta) = \sigma(a_i(\theta - b_i)) \quad (1)$$

where $\sigma(\cdot)$ denotes the logistic function. Intuitively:

- θ (**ability**) represents the overall skill level on the benchmark. Higher θ indicates a higher probability of answering difficult items correctly.
- b (**difficulty**) represents how challenging a question is. Items with larger b require higher ability to achieve 50% correctness (Embretson and Reise, 2013).
- a (**discrimination**) measures how sharply an item distinguishes between high- and low-ability respondents. Higher a means the item better separates strong from weak systems.

The raw parameters are not interpretable in isolation, but they enable comparison **between agents**. For example, a value such as $\theta = -2.91$ does not correspond to "-291% accuracy" or any direct percentage. Rather, it indicates that the model's ability is substantially below the dataset's average difficulty level (which is centered near 0). Concretely, when $\theta \ll b$, the logistic function yields a low probability of correctness across many items, especially those with moderate or high discrimination. Thus, strongly negative θ reflects consistent failure even on moderately difficult questions, not just lower raw accuracy.

The numbers reported in **Table 4** are obtained by fitting the 2PL model to the binary correctness matrix (respondent $\times$ item) via maximum likelihood estimation. Item parameters (a_i, b_i) and respondent abilities (θ_j) are estimated jointly under standard identifiability constraints (e.g., fixing the mean ability to 0). Once fitted, each model's θ is directly obtained from the estimated parameters.

B Data Collection and Processing Details

Quizmasters Website The Quizmasters website[2] publishes standalone audio clips grouped under categorical umbrellas, but without associated questions. These clips range from 3 to 40 seconds in length and are organized into categories that either apply audio transformations (e.g., reversal) or present challenging identification tasks involving less popular material. Since clips within a category share common properties, we assign handwritten questions appropriate to the category format. For example, a clip from the National Anthems category may be paired with the question *"What country is this national anthem from?"* Some collections also include a corresponding reveal clip containing the original, untransformed audio; because these clips do not exhibit the intended transformation, we instead assign identification-style questions such as "Name the artist who recorded this song." In addition to question–answer pairs, we store clip-level metadata including sampling rate, duration, and other audio attributes.

PAVEMENT Processing Each dataset entry includes a *notes* field that preserves additional information from the original QuizBowl questions, such as alternative acceptable answers or clarifying remarks. These annotations are retained as part of the source material. In the PAVEMENT subset, QuizBowl questions are split into individual clues for model evaluation.

Some questions refer to shared attributes across clues, such as *"What is the common number in the titles of these songs?"* or *"What is the common profession mentioned in the titles of these songs?"* While some tournaments, such as SoundTrack, provide clip descriptions in their GitHub repositories, this practice is inconsistent and absent in PAVEMENT.

Our processing of PAVEMENT focused on data integrity. Earlier scraped versions contained mismatches between audio clips and questions, ordering errors, and formatting issues. We therefore re-scraped the dataset to ensure correct question–answer alignment and applied normalization steps to make outputs human- and model-ready, including resolving Unicode issues, inconsistent notation, and formatting irregularities.

Data Preparation

We prepare AUDITA in three stages: (i) extraction and alignment, (ii) normalization for evaluation, and (iii) categorization for analysis. The goal is not to change the underlying questions, but to make the resulting audio–question–answer triples consistent, human-readable, and robust to evaluation artifacts.

Extraction and alignment. For GitHub-hosted QuizBowl-style sources, we extract question prompts and answerlines from the provided materials and align them with audio files using source-specific directory conventions and indexing. For

[2] https://www.thequizmasters.biz/

Pavements, we correct earlier indexing mismatches by enforcing consistent clip identifiers and alignment rules between prompts, answerlines, and audio files. Table 6 reports the resulting category distribution.

External benchmark preparation. To provide a concrete comparison point to prior audio question answering resources, we include questions from OpenAQA and ClothoAQA. OpenAQA is largely generated from captions and metadata as part of LTU's OpenAQA-5M pipeline, while ClothoAQA is crowdsourced, which offers a small human-written counterpoint to caption-derived questions (Gong et al., 2023b; Lipping et al., 2022). We do not rewrite the benchmark questions or answers to make them more human-friendly. Instead, we apply only minimal preprocessing needed for evaluation consistency.

For OpenAQA, we run the filtering utilities released with LTU to remove unanswerable or hallucinated question–answer pairs (Gong et al., 2023b). In our snapshot, this removes 18.65% of candidate items. We then sample proportionally across OpenAQA's constituent source datasets to preserve its original mixture.

Normalization for evaluation. Raw answerlines and prompts contain encoding noise and formatting conventions that can create spurious evaluation failures for both humans and models. We apply multi-stage cleaning that targets: *(1) Character normalization:* mapping diacritics and special symbols to ASCII equivalents and removing Unicode artifacts introduced by heterogeneous source encodings. *(2) Formatting normalization:* stripping QuizBowl markup, removing prompt instructions and bracketed editorial artifacts, and rewriting alternative acceptable answers into a consistent "A or B" form. *(3) Context-aware cleanup:* for cases where rule-based edits are insufficient, we use GPT-4o-mini to rewrite answers into a standardized, human-readable form while preserving semantic content and listed alternatives. When cleanup is uncertain, we prefer conservative edits that preserve the original label over aggressive rewriting.

GPT-4o-mini was used strictly for **format normalization and controlled answer canonicalization**, not for modifying semantic content or introducing new answer keys. Specifically, its role was limited to:

1. Removing encoding artifacts and markup (e.g., QuizBowl formatting such as underlines and bracketed instructions).

2. Converting acceptance instructions into explicit canonical surface forms.

For example, the original annotation:

> *""Pathétique" Sonata [or Beethoven's Piano Sonata No. 8 in C minor, Op. 13 (accept any underlined part)]"*

was normalized to:

> *"Pathetique Sonata or Beethoven's Piano Sonata No. 8 in C minor or Sonata No. 8 in C minor"*

No new aliases were added beyond those explicitly licensed by the original annotation ("accept any underlined part"), and no semantic reinterpretation was performed. The model was used only to rewrite formatting artifacts into clean, evaluation-ready strings while preserving the original acceptance set.

Statistic	Pavements	Audio-Packets	Quizmasters	External Sources
Questions (%)	6.94%	17.02%	42.70%	33.33%
Avg. Audio Duration (s)	63.42	65.25	41.81	13.75
Avg. Ques. Length	6.56	4.68	14.54	12.04

Table 5: Distribution of questions in the AUDITA Benchmark Dataset by source, with audio duration and question length statistics.

C OpenAQA Filtering and ClothoAQA Evaluation Setup

OpenAQA Filtering The 18.65% removal follows the filtering script released by the original OpenAQA authors (Gong et al., 2024). This script excludes GPT-generated responses that explicitly signal insufficient information, such as phrases like "cannot be determined," "not provided," "unclear," and related hedging expressions. Importantly, this is not a subjective quality-based filter; rather, it removes instances where the generative pipeline itself identified the question as unanswerable. The remaining 81.35% therefore consist of questions paired with committed, specific answers.

Positioning ClothoAQA Within Our Evaluation Framework ClothoAQA occupies a distinct position in our evaluation framework. Our critique focuses on its structural limitations as a standalone benchmark, including ambiguous questions, overlapping audio sources, and underspecified answer

Statistic	Who's Who? Name That Persona	Cultural Geography in Sound	Name The Music: Songs, Artists & Composers	Pop Culture and Media	Elements of Musical Works	Environmental and Acoustic Sound Recognition
% Questions	7.66%	4.56%	26.60%	19.13%	7.80%	34.24%
Avg. Audio Duration (s)	21.25	64.14	41.63	68.95	58.70	10.43
Avg. Question Length (words)	13.41	8.21	9.86	13.94	8.56	25.27

Table 6: Breakdown of questions in the AUDITA Benchmark Dataset by category, with average audio durations and question lengths.

Family	Typical Construction	Common Shortcut and What AUDITA Stresses	Typical Question Style & Example
Sound labeling	Closed label sets, short clips, event tags	Salient cue detection, weak long context. AUDITA uses longer clips and open answer space.	*Example:* Classify short environmental sounds (e.g., engine, applause).
Caption-derived QA	Questions derived from captions or metadata (e.g., OpenAQA (Gong et al., 2023b))	Lexical priors and caption artifacts. AUDITA uses human-authored trivia not derived from captions.	Caption-consistent attribute query. *Example:* Identify the sound source described by the caption.
Synthetic or templated	Generated scenes, fixed templates, constrained language	Template regularities and limited linguistic diversity. AUDITA uses natural, non-templated questions.	Programmatic logic over fixed attributes. *Example:* Count occurrences of a specified event.
Speech-centric QA	Spoken content dominates, transcript-like supervision	Can collapse to ASR plus text QA. AUDITA includes speech but also music and non-speech audio.	Spoken content identification, often transcript-based questions.
AUDITA	Human-written questions with real-world referents	Requires audio entity linking and multi-cue integration. This is the core target regime of AUDITA.	Probing trivia grounded in real referents. *Example:* Identify a film theme, speaker, or cultural artifact from audio cues.

Table 7: Positioning of AUDITA relative to common AQA benchmark designs, including typical dataset construction, common shortcuts that models exploit, and representative question styles and examples.

keys. At the same time, ClothoAQA is crowdsourced from human annotators rather than synthetically generated, placing it in a meaningfully different quality tier compared to caption-derived or templated datasets. Moreover, it was not included in OpenAQA, making it a genuinely independent data source. We include ClothoAQA to provide human performance baselines on a dataset already used by the community—not because we consider it free from the limitations discussed in Section 2.

Across the dataset, human evaluation collected 1517 human guesses-individual answer attempts by participants- providing a reliable set of judgments to benchmark model performance. In the next section, we describe our evaluation framework, which leverages these human guesses to jointly model question properties and participant abilities.

D Answer Formats

We evaluate two answer settings:

Free-Response Question Answering. In the free-response setting, models generate open-ended textual answers. Generated responses are evaluated using the PEDANTS (Li et al., 2024) framework, which determines semantic equivalence between model outputs and reference answers through structured normalization and equivalence rules, rather than relying on exact string matching.

Multiple-Choice Question Answering (MCQ). In the MCQ setting, models are given a fixed set of candidate answers and must select the correct option. We convert model outputs to a choice via either explicit option selection or scoring each candidate independently, depending on model capabilities. Accuracy is reported as the fraction of correctly answered questions.

E Evaluated Models: Architectures, Training Objectives, and Assumptions

We evaluate a diverse set of state-of-the-art open checkpoint models runnable locally, covering mid-scale language backbones (approximately 6B to 17B parameters in the language component) and a variety of audio front ends. These include models trained primarily on audio–text alignment objectives, large multimodal foundation models with audio inputs, and audio-specialized models adapted for question answering. To address the varied nature of questions—some requiring speech understanding such as lyrics or quoted lines—we categorize models into three groups: (i) audio-only understanding, (ii) speech-aware models, and (iii) unified audio+speech models, reporting results separately for each subgroup. For all models, we use publicly released checkpoints and follow recommended inference settings when available, with no fine-tuning on our dataset.

We evaluate a diverse collection of state-of-the-art open-checkpoint models designed for audio and multimodal language understanding, spanning mid-scale language backbones ranging from approximately 6 billion to 17 billion parameters. These models differ widely in their training objectives, architectural designs, and audio processing strategies. Among them are models primarily trained for audio–text alignment, such as Qwen2.5-Omni (Xu et al., 2025a) and AudioGPT (Huang

Dataset	Example Question	Answer (Gold)	IRT Difficulty / Discrimination	Issue / Notes
VGGSound QA (Chen et al., 2020)	*"What type of animal is making the high-pitched and sharp sound described in the audio?"*	*The high-pitched and sharp sound is most likely a bark produced by a dog.*	Low difficulty, low discrimination ($b = -2.1$, $a = 0.35$)	Simple audio classification task. Some clips have overlapping sounds causing ambiguity. Occasional metadata leakage possible, making some questions answerable without listening.
ClothoAQA (Drossos et al., 2020)	*"Is this outdoors?"*	*Yes*	Moderate difficulty, low discrimination ($b = -0.6$, $a = 0.42$)	Binary classification style question, with limited complexity. Contextual metadata sometimes gives away the answer. Not all questions require detailed auditory reasoning.
AudioCaps QA (Kim et al., 2019)	*"Create a brief audio description, create labels, caption next."*	*Labels: Vehicle; Tire squeal; Car; Race car, auto racing. Audio caption: Race car engines speed by, changing gears and screeching.*	Varied difficulty, often low discrimination ($b \approx 0.1$, $a = 0.28$)	Open-ended captioning questions. Subjective answers complicate evaluation and lack precise metrics. Do not constitute discrete QA.
"No listening needed" Questions (e.g. AudioCaps QA (Kim et al., 2019))	Examples: *"What is the most likely reason for someone to strike a metal trailer with a wooden rod?"* or *"What is the significance of thunder in mythology?"*	*There could be a variety of reasons, such as trying to get someone's attention, testing the durability of the trailer, or making a musical sound.* and *In many cultures, thunder represents the authority of gods and goddesses, and it is often associated with power, strength, and fertility.*	Very low difficulty, near-zero discrimination ($b = -3.1$, $a \approx 0$) and ($b = -3.4$, $a \approx 0$)	Answerable without listening, often appearing in scraped or poorly filtered datasets. Undermines auditory reasoning benchmarks. Not typical of curated datasets but important to highlight.

Table 8: Representative question examples from popular audio QA datasets, annotated with illustrative ranges of psychometric (IRT) difficulty and discrimination, and highlighting issues such as reliance on metadata, low reasoning complexity, and "no listening needed" questions. This underscores the need for carefully curated, human-authored datasets that robustly evaluate auditory reasoning.

et al., 2024), which focus on aligning acoustic features with language representations to handle speech and audio understanding tasks. In addition, we consider large multimodal foundation models that incorporate audio inputs alongside other modalities, including OpenOmni (Luo et al., 2025), Audio-Flamingo (Kong et al., 2024), and Phi-4-Multimodal (Abouelenin et al., 2025). These models leverage powerful language backbones integrated with audio encoders, enabling flexible multimodal reasoning across diverse input types. Specialized audio language models such as SALMONN-2 and SALMONN-2+ (Tang et al., 2025) emphasize improved audio–language alignment and training methods to enhance reasoning over complex auditory inputs, while models like MU-LLaMA (Liu et al., 2024) and the original SALMONN (Tang et al., 2024) also contribute to the landscape of advanced multimodal comprehension. Other models, including Qwen3-Omni (Xu et al., 2025b), LTU-AS (Gong et al., 2023a), Baichuan-Omni-1.5 (Li et al., 2025a), and VITA-1.5 (Fu et al., 2025), further expand the diversity of architectures and training approaches assessed.

For each model, we report performance on both text-based and multiple-choice (MCQ) question formats from our dataset, including accuracy percentages and item response theory (IRT) estimated ability scores (θ), which provide a latent measure of model proficiency relative to question difficulty. Models are ranked within each task format to facilitate comparative evaluation. All evaluations utilize publicly released checkpoints with recommended inference settings, without any task-specific fine-tuning, ensuring an unbiased benchmarking environment. This comprehensive evaluation enables us to analyze strengths and limitations across modalities, task types, and audio reasoning capabilities, thereby offering insights into current progress and challenges in the field of audio question answering.

F Human Evaluation and Participant Instructions

Participant Summary

- **Data collected:** 1517 individual answer attempts from 87 participants recruited from trivia communities.
- **Participant background:** Participants were not pre-selected experts but were generally familiar with common trivia domains (e.g., music, film, and pop culture).
- **Dataset coverage:** The dataset includes questions across multiple categories, allowing participants to engage with questions aligned with their strengths. This ensures that maximum achievable accuracy is not constrained by any single individual's expertise.

Issue	Dataset(s)	Example Question & Answer	Notes
Ambiguity	Clotho-AQA (Drossos et al., 2020), VGGSound QA (Chen et al., 2020), AudioCaps QA (Kim et al., 2019), FSD50K QA (Fonseca et al., 2022)	Q: *"What animal is making the sound?"* A: *bird* / *dog* (multiple audible) Q: *"Where is the sound coming from?"* A: *indoors* / *outside* (disagreement among participants) Q: *"What is the dominant sound in the clip?"* A: *siren* / *car horn* (subjective) Q: *"Is there a sound of laughter?"* A: *yes* / *no* (faintness ambiguity) Q: *"Is there a vehicle sound?"* A: *yes* / *no* (ambiguous engine/horn sounds)	Multiple overlapping sounds cause unclear targets; vague or underspecified question wording leads to inconsistent answers across annotators and models.
Weak Grounding / Shortcut	AudioCaps QA (Kim et al., 2019), AudioSet QA (Gemmeke et al., 2017)	Q: *"Is someone laughing?"* A: *yes* Q: *"Is there music playing?"* A: *yes*	Questions answerable from captions or metadata alone without listening; templated language encourages shortcut learning.
Underspecified Answer Keys	MUSIC-21 QA (Christodoulou et al., 2025), Clotho-AQA (Drossos et al., 2020)	Q: *"What instrument is playing?"* A: *piano* Q: *"What animal can be heard?"* A: *dog* / *puppy* / *hound*	Multiple valid lexical variants treated as separate answers, increasing noise.
False Presuppositions	Speech Commands QA (Warden, 2018)	Q: *"Is the command 'stop' present?"* A: *no*	Questions assume presence of commands that may not exist, confusing annotators and models.
Overlapping Audio	VGGSound QA (Chen et al., 2020), AudioCaps QA (Kim et al., 2019)	Q: *"What animal is making the sound?"* A: *dog*	Overlapping sound sources cause ambiguity, complicating correct labeling and answering.
Synthetic Question Bias	AudioSet QA (Gemmeke et al., 2017)	Q: *"Is there music playing?"* A: *yes*	Automated templated generation reduces linguistic diversity and causes models to exploit shortcuts.

Table 9: Summary of common issues in audio QA datasets, including improved, concrete ambiguous question examples from real datasets.

Overall, this evaluation provides a **robust estimate of human performance**, capturing both question difficulty and variability in participant knowledge.

We will provide detailed participant instructions in the Appendix and plan to release the collected human answers after blind review, supporting reproducibility and transparency.

Participants were also asked to select a category at the beginning of the task to ensure they answered questions within a domain they felt most comfortable with.

Instructions Given to Participants

1. **Two chances to answer:**
   - **First chance:** Type your answer freely in the text box provided.
   - **Second chance:** Select one answer from the four multiple-choice options.
2. **Feedback page:** After completing both responses, participants see a feedback page showing:
   - The correct answer
   - Their text response
   - Their multiple-choice (MCQ) selection
   - Whether the text response and MCQ choice were correct or incorrect
3. **Performance tracking:**
   - **Text Score:** Number of correct text responses
   - **MCQ Score:** Number of correct multiple-choice responses
   - Both scores are displayed at the end of the activity
4. **Audio timing:**
   - Automatically tracked to record how long each audio clip is played
   - Does **not** affect the final score
5. **Post-evaluation:**
   - All text responses undergo human review to ensure fair and accurate scoring

Tip: Listen carefully, use headphones if possible, and try your best on both attempts.

F.1 Participant List

We thank all participants for their contributions to data collection and evaluation. All participants provided consent to be acknowledged (Table 11).

G MCQ Analysis, Human Agreement, and Task Difficulty

MCQ Output Class Distribution Analysis We analyzed the class distribution of model predictions across all four answer options to better understand convergence behavior and potential bias in the MCQ setting.

Aggregating across all models and items, the output distribution is shown in Table 13. Importantly, the model does not collapse onto a single dominant option, and predictions are distributed across all

Text Questions				
Dataset	Modality	Acc (%)	Mean θ	SD σ
Pavements	Human	35.28	0.09	1.74
Pavements	Model	4.26	-2.81	0.60
Audio-Packets	Human	34.24	0.08	0.61
Audio-Packets	Model	8.89	-2.03	0.67
Quizmasters	Human	21.34	0.03	0.91
Quizmasters	Model	1.58	-3.88	0.62
External Datasets	Human	25.79	0.05	0.51
External Datasets	Model	13.49	-1.45	0.51
MCQ Questions				
Dataset	Modality	Acc (%)	Mean θ	SD σ
Pavements	Human	53.90	0.11	1.75
Pavements	Model	5.89	-2.33	0.61
Audio-Packets	Human	61.21	0.13	0.62
Audio-Packets	Model	11.76	-1.83	0.65
Quizmasters	Human	50.31	0.10	0.93
Quizmasters	Model	3.04	-2.70	0.62
External Datasets	Human	74.30	0.15	0.49
External Datasets	Model	20.59	-1.11	0.50

Table 10: Comparison of accuracy and mean ability (θ) across datasets, modalities (human vs. model), and question types (Text vs. MCQ) shows humans consistently outperform models. Pavements and Audio-Packets yield moderate human accuracy (≈30–60%) with positive abilities, while models score lower with negative abilities. Quizmasters is more challenging, especially for models. External datasets show the highest human accuracy, notably on MCQs. These results reveal clear human-model gaps and varying task difficulty by dataset and question type.

four classes. The distribution remains relatively balanced overall. This suggests that the below-chance MCQ accuracy (15.65%) is not driven by trivial positional collapse or degenerate convergence behavior.

Instead, the pattern is more consistent with systematic selection of semantically plausible but incorrect distractors. In other words, the model is neither guessing uniformly at random nor defaulting to a single class; rather, it appears to rely on partial cues that lead to structured but incorrect choices (Table 13).

This analysis highlights that errors are distributed fairly evenly across distractors, further supporting that model failures arise from structured but incorrect semantic choices rather than simple positional bias.

Inter-Human Agreement Questions were answered by multiple participants, enabling aggregation of correct answers across individuals to approximate a reliable human topline.

Average agreement across participants varies by question type but demonstrates that questions are answerable given attention and domain knowledge. Inter-human agreement (top 20% skilled participants, PEDANTS-normalized exact match) is 0.91 for open-ended questions (1.0 in MCQ format), indicating very high convergence among the most consistent participants despite the open-ended format.

This supports that the questions are well-posed and answerable, while remaining challenging.

Human Topline Score and Task Difficulty The human topline score of 32% reflects the difficulty of open-ended audio reasoning, not infeasibility. All questions are collected from real-world trivia competitions, curated online tournaments, and expert-designed quizzes (Section 3).

These questions are explicitly intended to be answerable by humans, with each item verified and crafted to ensure solvability given attentive listening and relevant knowledge. For example, participants might be asked to identify a composer from a musical motif or an actor speaking in a clip.

Such tasks require careful auditory perception combined with cultural or domain knowledge. The 32% score highlights the intrinsic challenge of precise auditory reasoning in open-ended text responses, rather than a lack of information (Table 12).

Explanation for Below-Chance MCQ Performance To better understand model behavior on multiple-choice questions, we analyzed prediction patterns, human agreement, and task difficulty (Appendix G for full tables and examples). Models achieve an average MCQ accuracy of 15.65%, below the nominal 25% chance level for four-option questions. This below-chance performance is not caused by dataset flaws or trivial positional collapse; instead, it arises from structured confusion among plausible distractors, reliance on subtle audio cues, and the inherently challenging nature of open-ended audio trivia. For instance, in the question *"Name the title character of these movies"* (correct answer: *Batman*), the model selected *Superman* among highly plausible alternatives, illustrating systematic but semantically reasonable errors.

Analysis of class distributions confirms that predictions are spread across all four options rather than collapsing to a single choice (see Table 13 in Appendix G). Similarly, error distribution across distractors is relatively balanced, supporting the interpretation that models make structured misselections rather than random guesses.

Participant Name						
Daniel Kim Stefany Meyer	Sara DelVillano Jason Christopher	Raymond Kimball Forrest Weintraub	Drew Scheeler Nishant Balepur	Nathan Zhao Eve Nuria Fleisig	Mohammed Afaan Mohammed Arif Ansari	Hemanth Nandakumar

Table 11: List of participants who contributed to this study and agreed to be acknowledged by name.

Human validation further demonstrates that these questions are answerable: the top 20% of participants achieve near-perfect agreement (1.0 in MCQ format), and the open-ended human topline reaches 32%, highlighting the intrinsic difficulty of precise auditory reasoning rather than dataset ambiguity. Representative high- and low-difficulty examples are provided in Appendix G.

Taken together, these findings indicate that AUDITA questions are challenging yet solvable, and that model failures primarily reflect the interplay of perceptual, semantic, and distractor-driven challenges rather than flaws in dataset construction.

The observed MCQ accuracy of 15.65%, which is below the chance level of 25% for 4-choice questions, arises from the combination of extremely large answer spaces in the open-ended audio trivia domain and the way distractors are designed.

1. Systematic confusion: Many questions contain plausible distractors that are similar to the correct answer, leading models to select incorrect options more frequently than random guessing.

2. Question–audio alignment: Some questions rely on subtle audio cues that the model cannot reliably perceive, further skewing predictions toward incorrect distractors.

3. High difficulty nature: The dataset contains rare or niche knowledge (e.g., obscure songs, actors, or composers), which is challenging even for models with strong general language understanding, resulting in lower-than-chance accuracy in multiple-choice settings.

Thus, the below-chance performance does not indicate a quality issue with the dataset, but rather reflects the combination of carefully constructed distractors, high question difficulty, and reliance on perceptual audio understanding, which models currently struggle to handle.

For example, consider the question: *"Name the title character of these movies."* The correct answer was *Batman* (acceptable variants included *Bruce Wayne* or *Dark Knight*), with distractors *Spider-Man*, *Superman*, and *Iron Man*. The model selected

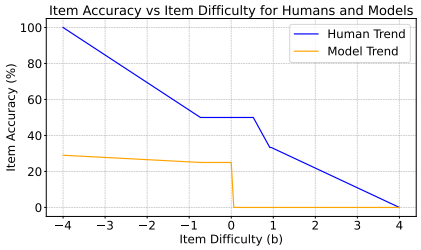

Figure 4: Item accuracy plotted against item difficulty (b) for humans (blue) and models (orange). Humans generally show higher accuracy than models across difficulties; however, some human accuracy values drop to zero due to sparsity of responses on certain items. The trend lines, highlight the widening performance gap between humans and models.

Superman. All options are highly plausible superhero characters, and without confidently grounding the audio cues, the model gravitates toward semantically related but incorrect alternatives.

When such confusion occurs systematically across many items with strong distractors, accuracy can fall below the nominal 25% chance level for four-option MCQs—not because of dataset flaws, but because models exhibit structured miscalibration rather than random guessing.

For reference, the top 20% of human participants achieve near-perfect agreement, with MCQ responses scoring 1.0 (agreement) and an overall accuracy of 89.33%. This demonstrates that, while models struggle below chance, high-performing humans are able to reliably answer the same questions, confirming that the dataset is answerable and that the low model performance reflects task difficulty and distractor design rather than dataset flaws.

Questions are Challenging but Gettable When a set of questions exhibits low accuracy, it is natural to suspect that the dataset may contain false presuppositions, ambiguities, or incorrect answer keys. One of the goals of the human validation process is to ensure that this is not the case for this dataset.

Item Response Theory (IRT) analysis provides additional verification. Most examples are answered correctly by high-skilled human participants, with only 26 examples never receiving a correct human response. Average agreement across

participants varies by question type but demonstrates that questions are answerable given sufficient attention and domain knowledge.

Inter-human agreement (PEDANTS-normalized exact match) is 0.36 (0.57 in MCQ format), indicating substantial convergence among participants despite the open-ended format. These results support that the questions are well-posed and answerable, while remaining challenging.

H Related Work

Audio question answering (Audio QA) Audio Question Answering remains a nascent field with relatively few datasets, many of which impose design constraints that limit their ability to evaluate genuine auditory reasoning. Early efforts such as CLEAR (Lin et al., 2021) construct synthetic acoustic scenes by layering individual musical notes from the GoodSounds database. Questions are programmatically generated from logical templates and target specific attributes, for example, *"How many times does the C note occur in this clip?"* While this approach enables precise semantic control, it restricts linguistic diversity and limits reasoning complexity.

Similarly, DAQA (Fayek and Johnson, 2020) composes variable-length audio clips from a closed vocabulary of 32 sound classes (e.g., ansdog bark, *car horn*) and asks questions such as *"Does the sound of a car horn occur more than twice in this clip?"* Although DAQA allows limited temporal reasoning, its answer space is restricted to yes/no or counts, constraining the evaluation of richer auditory inference.

More naturalistic datasets attempt to move beyond synthetic audio. ClothoAQA (Drossos et al., 2020) relies on crowd workers to write questions about environmental sound recordings originally collected for captioning. Questions such as *"What animal makes the sound in this clip?"* or *"Is the sound recorded indoors or outdoors?"* better resemble real-world queries. However, this process introduces strong linguistic priors: models trained on ClothoAQA perform competitively even when audio is removed, indicating reliance on textual cues rather than acoustic understanding. MusicAVQA (Li et al., 2022) exhibits similar limitations, using templates like *"What instrument is playing?"* or *"Is the tempo fast or slow?"* These formats further limit linguistic variability and encourage shortcut learning, echoing issues observed in textual and visual QA (Section H).

Audio Question Answering Benchmarks Existing AQA benchmarks provide important testbeds for audio–language modeling but do not systematically expose failures in human-relevant auditory reasoning. CLEAR and ClothoAQA primarily reduce to controlled attribute queries or implicit classification tasks, making it difficult to distinguish true reasoning from surface-level pattern matching. While foundational, these benchmarks are limited in their ability to reveal nuanced model weaknesses in realistic, probing scenarios.

More recently, the DCASE 2025 Audio Question Answering challenge (Yang et al., 2025) introduced multi-domain QA subsets spanning bioacoustics, temporal soundscapes, and complex real-world audio. Although this effort broadens domain coverage, it remains centered on multiple-choice evaluation and lacks an adversarial, human-authored component designed to surface model brittleness.

Complementary evidence from Wang et al. (2025a) shows that large audio–language models are highly sensitive to misleading or conflicting textual cues paired with audio, revealing robustness gaps in current evaluations. Together, these findings motivate benchmarks that deliberately incorporate adversarial examples and human-authored questions to stress robust audio–language reasoning.

Adversarial Evaluation in QA and Multimodal Tasks Outside of audio QA specifically, adversarial evaluation has been successfully applied in other QA domains. In multimodal QA such as Visual Question Answering, human-in-the-loop adversarial data collection (e.g., Adversarial VQA) has been shown to produce questions that systematically expose model weaknesses by allowing annotators to target model failure modes through iterative feedback. Studies on adversarial QA in text also reveal that without adversarial examples, models can achieve high accuracy by exploiting dataset biases rather than robust reasoning. These findings underscore the value of adversarially collected questions for diagnosing model behavior, but comparable efforts are scarce for purely auditory content.

To address these gaps, we construct a human-authored adversarial audio QA dataset grounded in real audio recordings and designed to be easy for humans but challenging for current models. Unlike prior AQA benchmarks, our dataset emphasizes semantic richness, adversarial focus, and natural

realism, enabling evaluations that reveal model brittleness that structured or synthetic datasets fail to surface. **Multimodal Question Answering** One of the first large-scale multimodal datasets for question answering was the VQA dataset (Antol et al., 2015), which used image data to give models the context to answer a natural language question. The authors had crowd workers write questions about images from the COCO dataset and synthetically generated scenes to produce the data. Shortly after VQA was released in 2015, Johnson et al. (2017a) of the CLEVR dataset created a VQA task that uses fully synthetic scenes to produce a comprehensive visual reasoning test (Johnson et al., 2017b). One of the key aspects of this dataset was that questions were generated using a functional program, which would inspire future Visual Datasets.

One of the main issues within the VQA task was the presence of heavy priors within the data, where emergent statistical patterns would undermine the goal of reasoning over the image and the text. Many groups have been making efforts to improve upon this. The VQA v2 dataset balances the VQA dataset by introducing complements to each data point, where the new image is similar to the original but produces a different answer to the corresponding question (Goyal et al., 2017). Later, efforts were made to control the distribution of answers by (Agrawal et al., 2018) of the VQA-CP, who changed the splits of the VQA and VQA v2 datasets to alter the priors of the answer distribution. While remedies have been made to the original VQA dataset to fix its issues with heavy bias, other datasets have been introduced to overcome these pitfalls. One such dataset is GQA, which uses scene graphs based on images from COCO and Flickr to build questions automatically (Hudson and Manning, 2019). From the scene graph, questions and answers are built from a functional program similar to what was used in CLEVR, which further allowed the authors to smooth the answer distribution for various groups of questions.

Outside of entirely image/text-based multimodal QA datasets, several examples of datasets explore different mediums and combine already popular ones. For example, the MultimodalQA creates multimodal questions by composing single modality questions about Wikipedia tables and the entities linked within them (such as images or other objects) (Talmor et al., 2021). To compensate for the algorithmic generation of questions, the authors use crowd workers to rephrase the question into a more natural alternative and have other workers verify the question's validity. Several video-based datasets have also been released following the pattern of looking at combinations of modalities. For example, Yang et al. (2022) used videos from the VGG-sound dataset and expert annotators to write questions about each video for AVQA dataset. Similarly, Li et al. (2022) released the Music-AVQA dataset by collecting YouTube videos of music performances and crowd workers produce questions that followed a predefined template.

While many of these datasets utilize datasets from adjacent tasks of similar modality, only a subset of those available are web-curated. An advantage of our dataset in this field is that humans have already written the questions we collect outside of the context of our research, which minimizes much of the bias observed from directly using crowd workers to produce data for a benchmark.

Audio Question Answering – AQA datasets have been generally sparse over the past 6 years. Major contributions were either synthetically generated like DAQA (Fayek and Johnson, 2020) and CLEAR (Lin et al., 2021) or reliant on crowd workers for annotations like ClothoAQA (Drossos et al., 2020). One of the earliest examples of a contemporary AQA dataset is the CLEAR dataset, which shares many similarities with the CLEVR dataset. In particular, the CLEAR datasets combines individual musical notes from the GoodSounds database to generate an acoustic scene for a model to analyze. Like CLEVR, CLEAR's questions are constructed from templates represented as a logical tree with a functional program associated with them (Lin et al., 2021).

Using a similar methodology of combining smaller events, the DAQA dataset constructs audio-question pairs by stitching together several audio events. The main difference between the two is that DAQA uses events of variable length at various frequencies, so they can ask questions about how often a specific event occurs within a clip. While this dataset can help test a surface level of reasoning, the answer space is very small, with only 32 classes, with many answers being yes or no. Later, the Clotho AQA dataset was made, which takes advantage of the Clotho audio captioning dataset and uses crowd workers to produce new questions. Regarding crowd work, the resulting dataset may often contain heavy priors as quality control has been difficult. There have been several attempts to treat this issue in the VQA space (Anderson et al.,

2017); however, this hasn't been extended to the world of audio yet. Notably, this issue also skews results from this dataset, as many of the experiments conducted simply answering the question yield higher performance than using the audio and the question.

Adversarial Dataset Creation – As models grow increasingly complex, it becomes significantly more difficult to understand precisely why a model makes particular decisions during its inferences. A consequence of this trend is that understanding the weaknesses of a model turns into a hard task, especially when it comes to black-box models. The goal of adversarial dataset generation is to explore how robust models are to noisy data and to look at ways models underperform compared to humans. Thus, an important quality of adversarial questions is that they are hard for computers but relatively easy for humans; if it is very difficult for both groups, then it simply shows the dataset may be too tough. For VQA, a group of researchers made Adversarial VQA (AVQA), which uses a human-in-the-loop approach to generate questions that challenge models. An important aspect of their question creation interface is that a SOTA VQA model is present to provide answer feedback, so people can tweak questions until the model is tricked (Sheng et al., 2021). After the questions have been written, the model is retrained and given to the workers to write new, more difficult questions. In an interesting case of coincidence, within the same year, Li et al. (2021) produced a dataset called AdVQA using a similar technique. Although their annotation processes differ, AdVQA doesn't use model retraining. An interesting example of intentionally designing a challenging dataset is TriviaQA, a reading comprehension dataset that grounds its queries in trivia questions (Joshi et al., 2017). Since Trivia questions are designed to test the ability of humans, they are also a great way to benchmark models, as they are written to find out the best of a group of people.

Item Response Theory in NLP evaluation Item Response Theory (IRT), originally developed in psychometrics, has recently gained traction in NLP as a framework for evaluating both datasets and models. Early work by Lalor et al. (2016) introduced IRT to NLP by constructing evaluation sets for Recognizing Textual Entailment (RTE), demonstrating that standard metrics such as accuracy assume all items are equally informative, whereas IRT models item-specific properties such as difficulty and discrimination. This allows evaluation to account for heterogeneity in question quality and provides a more nuanced measure of system performance relative to human populations.

Subsequent work has expanded IRT beyond dataset construction to broader evaluation and analysis. Lalor (2020) show that IRT parameters can be estimated using model-generated responses ("artificial crowds"), enabling scalable application without extensive human annotation. More recent efforts apply IRT to diagnose biases and failure modes in NLP systems, using item parameters to quantify how model performance varies across demographic or linguistic factors (Xu et al., 2025c).

Recent studies further highlight the role of IRT in rethinking benchmark evaluation for large language models. For example, Zhou et al. (2026) argue that traditional leaderboards often fail to distinguish top-performing models due to poorly calibrated or low-discrimination items, and propose IRT-based frameworks to improve benchmark reliability and separability. In parallel, tutorials and surveys (e.g., Lalor et al. (2024)) emphasize the growing adoption of IRT as a general tool for analyzing dataset quality, estimating model ability (θ), and identifying informative versus uninformative examples.

Overall, this line of work shows that IRT provides a principled alternative to aggregate metrics by jointly modeling item characteristics and system ability, enabling more reliable comparison, diagnostic analysis, and benchmark design.

I Result

I.1 Model Scale and Validity of Conclusions

Many audio-language models couple a pretrained audio encoder to a pretrained text LLM by projecting audio features into the LLM token space, so the language backbone is a key determinant of downstream instruction-following and knowledge-heavy QA performance (Gong et al., 2023b). Evidence from recent scaling analyses of speech-text and audio-centric language models also supports the general trend that larger backbones and stronger decoders improve aggregate performance across understanding and reasoning tasks, although gains vary by task and domain (Cuervo and Marxer, 2024).

At the same time, scale alone does not necessarily resolve failures caused by weak audio ground-

Difficulty	Question	Answer
High (Low feasible)	What country is this national anthem from?	Bolivia
Low (High feasible)	What is the language spoken in this clip?	Chinese

Table 12: Example questions illustrating task difficulty.

Option	Prediction Distribution (%)	Error Distribution (%)
Option 1	25.45	24.90
Option 2	29.16	28.92
Option 3	24.65	24.52
Option 4	20.74	21.66

Table 13: Model output distribution across MCQ answer options compared with distractor-level error distribution aggregated across all models. Errors are distributed fairly evenly across options, suggesting that below-chance performance is not caused by positional collapse.

ing and over-reliance on textual priors, which can manifest as confident but incorrect answers when audio evidence is insufficient or ignored (Wang et al., 2025b). This is consistent with MMAU, which reports that even strong proprietary systems remain far below human performance on its human-evaluated test-mini split and are only moderately separated from strong open models on the benchmark evaluation (Sakshi et al., 2024). For example, MMAU reports human performance of about 82.23 on test-mini, while Gemini Pro v1.5 achieves 52.97 on the test split and strong open models such as Qwen2-Audio-Instruct are comparable in several settings (Sakshi et al., 2024). These gains are meaningful, but they are not of a magnitude that would plausibly convert near-chance behavior on a difficult, human-authored benchmark into reliable audio reasoning.

Accordingly, while our evaluation focuses on open models in the mid-scale regime, our qualitative conclusions about failure modes and dataset difficulty are unlikely to be artifacts of model scale alone. Exhaustively evaluating cloud-scale proprietary models across all 9,690 questions would also impose substantial cost, which limits full scaling sweeps in this study.

J Understanding Model Failures Across Modalities

J.1 Factual Knowledge vs. Audio Understanding

We agree that certain questions (e.g., *"Name the artist who recorded this song"* or *"What film is this theme from?"*) require both perceptual recognition and world knowledge. To directly address this concern, we conducted additional analyses leveraging two model settings already available in our experiments: Question + Transcript (Whisper output) and Question + Raw Audio.

This comparison allows us to disentangle whether performance differences stem from: Lack of perceptual audio understanding, Imperfect transcription, Insufficient factual/world knowledge. Specifically: If a model succeeds with the transcript but fails with the audio, this indicates perceptual audio processing limitations. If it fails in both settings, the bottleneck is more likely knowledge-based. If it succeeds with audio but not transcript, this suggests information present in non-verbal acoustic cues (e.g., melody, timbre, instrumentation) that transcripts cannot capture.

Importantly, accuracy is substantially higher in the raw-audio condition (e.g., for GPT-4o 14.87%) than in the transcript condition (e.g., for GPT-4o 7.26%). If performance were primarily driven by factual recall from textual cues, we would expect transcript-based performance to match or exceed audio-based performance. Instead, the lower transcript accuracy suggests that automatic transcriptions omit or distort critical acoustic information necessary for correct identification.

Our results show that performance gaps persist even when transcripts are provided, indicating that failures cannot be attributed solely to missing encyclopedic knowledge. Moreover, the gap between transcript and raw-audio conditions highlights genuine audio-understanding limitations, not merely factual recall deficiencies (Table 18).

Performance gaps persist even when transcripts are provided, indicating that failures cannot be attributed solely to missing factual or world knowledge. The gap between transcript and raw-audio conditions highlights genuine audio-understanding limitations, not merely factual recall deficiencies. These examples illustrate how the two modalities (transcript vs. audio) contribute differently, enabling classification of failures as perceptual, knowledge-based, or audio-cue-driven.

J.2 Failure cases analysis

We conducted a systematic error analysis over all model errors and categorized them using the same diagnostic logic described in the rebuttal (audio succeeds / transcript fails; transcript succeeds / audio fails; both fail). The breakdown over error cases is as follows: Knowledge-based errors (both transcript and audio fail): 78.23%, Perceptual errors (transcript succeeds, audio fails): 8.82%,

Top 10 Best Discriminator Questions (IRT)		
#	**Question**	**Answer**
1	Name the title character of these movies.	*Sherlock Holmes*
2	What is the name of the person who is speaking in this clip?	*Harry Styles*
3	Name the character that inspired this music.	*Batman*
4	Name the lead artist.	*Halsey*
5	Give the common word found in the names of these lead artists.	*A$AP*
6	Name the artist.	*beabadoobee*
7	What country is this national anthem from?	*Bolivia*
8	What TV show is this clip from?	*A Team*
9	What TV show is this clip from?	*Happy days*
10	Name the city where these movies are wholly or mostly set	*Paris*

Top 10 Worst Discriminator Questions (IRT)		
#	**Question**	**Answer**
1	What is the language spoken in this clip?	*Chinese*
2	What is the language spoken in this clip?	*German*
3	What is the next line of lyrics that occurs after the song in the clip ends?	*Two and Two Were Four*
4	What is the next line of lyrics that occurs after the song in the clip ends?	*On the Pages in Between*
5	What country is this national anthem from?	*Iceland*
6	Name the character.	*Siegfried*
7	Name the mythical figure who is singing in these excerpts.	*Hades*
8	Name the male lead of these movies	*Charlie Chaplin*
9	What type of role do these characters have in common?	*Trouser role or Pants role or Breeches role or Male roles played by women*
10	What is the name of the person who is speaking in this clip?	*Bruce Forsyth*

Table 14: Examples of the top 10 best and worst discriminator questions by Item Response Theory (IRT), including their answers. High discrimination indicates questions that effectively differentiate between high- and low-ability respondents, while low discrimination indicates poor differentiation power.

Audio-cue errors (audio succeeds, transcript fails): 12.95%. Several clarifications are important for interpreting these numbers.

First, approximately 21.77% of errors (8.82% + 12.95%) are directly attributable to audio-related limitations. These cases are not reducible to missing encyclopedic knowledge. Second, the predominance of knowledge-based failures is consistent with the overall low accuracies (e.g., 14.87% audio; 7.26% transcript for GPT-4o). Many questions require identifying specific composers, performers, fictional characters, or cultural artifacts; even perfect perception would not suffice without the relevant world knowledge.

However, a substantial subset of transcript failures arises because the automatic transcription is empty, fragmentary, or effectively gibberish when the audio contains no linguistic content (e.g., instrumental music, sound effects, or non-verbal acoustic signatures). In these cases, the transcript condition is structurally disadvantaged: failure cannot be attributed purely to missing world knowledge, since the relevant signal exists only in the acoustic domain.

Representative examples are shown in Table 19.

In each case, the transcript contains either no usable information or severely degraded content. The identifying signal (melody, orchestration, anthem structure) exists only in the acoustic domain. Thus, these failures reflect structural information loss in the transcript condition rather than purely missing world knowledge.

Overall, while knowledge-based failures account for 78.23% of errors under our categorization scheme, this figure should not be interpreted as indicating that these failures are *purely* knowledge-driven. The categories are analytically defined for diagnostic clarity, but in practice, many items involve overlapping bottlenecks.

In particular, a substantial subset of cases classified as "knowledge errors" occur in settings where the transcript is empty, fragmentary, or unintelligible due to the absence of linguistic content (e.g., instrumental passages, orchestral themes, non-verbal acoustic signals). In such cases, the transcript condition contains little or no recoverable evidence. While successful identification still ultimately requires world knowledge (e.g., recognizing a specific anthem or opera), the failure cannot be attributed to missing knowledge alone—the relevant

perceptual signal is either unavailable or degraded in the transcript representation.

Thus, the 78.23% figure reflects the *final failure mode* (i.e., the model did not retrieve the correct entity), but many of these items are jointly constrained by:

- Knowledge requirements (e.g., knowing the composer, work, anthem, or character), and
- Audio-specific limitations or structural signal loss (e.g., non-linguistic music that transcripts cannot encode).

Importantly, 21.77% of errors are explicitly modality-driven (8.82% perceptual; 12.95% audio-cue), and an additional portion of the "knowledge" category includes items where transcript degradation meaningfully contributes to failure. Therefore, although knowledge is the largest single labeled category, the empirical picture is not one of purely encyclopedic deficiency. Rather, failures frequently arise from the interaction between knowledge demands and modality-specific constraints.

Model	LLM Core	Inputs	Outputs	Rationale
Omnimodal models (6)				
Qwen2.5-Omni	7B	text, audio, image, video	text, speech	Thinker–Talker architecture with explicit reasoning–speech decoupling and unified multimodal support.
Qwen3-Omni	30B-A3B	text, audio, image, video	text, speech	MoE-based omni model (30B total, $\sim$3B active) optimized for scalable multimodal reasoning and streaming speech.
OpenOmni	7B	text, audio, image	text, speech	Language-pivot alignment with progressive modality training and preference-based speech tuning.
VITA-1.5	7B	text, audio, image, video	text, speech	End-to-end omni model using a three-stage training pipeline without cascaded ASR/TTS.
Mini-Omni2	0.5B	text, audio, image	text, speech	Lightweight end-to-end omni assistant with parallel text–audio decoding and command-based duplex interruption.
Baichuan-Omni-1.5	7B	text, audio, image, video	text, speech	Unified decoder with explicit audio token modeling and staged multimodal training, optimized for Chinese–English bilingual use.
Audio-language models (4)				
Audio-Flamingo	1.3B	text, audio	text	Flamingo-style gated cross-attention, sliding-window audio features, ICL/RAG and multi-turn dialogue.
Qwen2-Audio	7B	text, audio	text	Whisper-large-v3–initialized audio encoder into Qwen-7B, unified prompting, SFT+DPO alignment.
LTU-AS	7B	audio, text	text	Frozen Whisper perception + TLTR time/layer aggregation, continuous audio tokens + transcript, LLaMA-7B w/ LoRA.
MU-LLaMA	7B	text, audio	text	Frozen MERT music encoder + adapter injection into LLaMA-2 7B, MusicQA supervision for QA and captioning.

Table 15: Evaluated models organized by capability grouping (Part 1 of 2). All models evaluated in audio-question to text-answer setting.

Model	LLM Core	Inputs	Outputs	Rationale
Speech-capable models (6)				
Phi-4-Multimodal	3.8B	text, audio, image, video	text	Mixture-of-LoRAs with frozen language backbone, modality-specific adapters, 128K context.
SpeechGPT	13B	text, speech	text, speech	Discrete speech units expanded into LLaMA vocabulary, three-stage training with Chain-of-Modality instruction tuning.
AudioGPT	modular	text, speech, audio, music, image	text, audio, video	Modular orchestration system using ChatGPT to coordinate 16+ foundation models via ASR/TTS interface.
SALMONN	13B	text, speech, audio, music	text	Dual encoder (Whisper + BEATs), window-level Q-Former, activation tuning for emergent abilities.
video-SALMONN 2	7B	text, audio, video	text	Frozen backbone with audio branch, MrDPO for caption optimization, atomic event-based quality metrics.
video-SALMONN 2+	7B	text, audio, video	text	Caption-enhanced training via MrDPO-generated data, SOTA on Video-MME/WorldSense/AVUT benchmarks.

Table 16: Evaluated models organized by capability grouping (Part 2 of 2). All models evaluated in audio-question to text-answer setting.

Dataset / Paper	**Example or Description**	**Why Models Excel**
VGGSound (Chen et al., 2020)	Models detect synthetic or repeated alarm/beep sounds perfectly. These sounds have highly structured, repetitive waveforms easy for pattern matching.	Models trained on millions of audio clips memorize these patterns and detect them with near-perfect accuracy, often better than non-expert humans.
Speech Command Recognition (benchmark dataset) (Warden, 2018)	Recognizing isolated spoken command keywords like "stop," "go," or "yes" — models achieve >99% accuracy, often exceeding average human recognition.	Limited vocabulary and clean synthetic data make these tasks trivial for models.
AudioSet Tagging (Gemmeke et al., 2017)	Models detect environmental sounds like sirens, horns, or machine noises with very high precision, sometimes outperforming humans in noisy clips.	Large training data and strong feature extraction enable models to spot subtle acoustic cues missed by humans.

Table 17: Examples of tasks where models demonstrate superhuman or near-superhuman performance in audio question answering or classification.

Case	Question	Answer	Transcript Result	Audio Result	Explanation
Transcript succeeds, audio fails (perceptual limitation)	Name the songwriter	Joni Mitchell (*River*)	✓	×	Transcript contains lines from the song, allowing the correct answer. Raw audio alone does not provide enough cues for the model to identify the song.
Audio succeeds, transcript fails (acoustic cue reliance)	Name the composer	Carl Orff	×	✓	Distinctive choruses and instrumental timbres in the audio enable the model to answer correctly; the transcript does not contain useful information.
Fails in both (knowledge limitation)	You are listening to the voice of a fictional character; what is this character's name (do not give the name of the actor/actress)?	Frank Spencer	×	×	Requires encyclopedic knowledge of TV characters; neither transcript nor audio alone is sufficient.

Table 18: Examples illustrating how transcript and raw-audio conditions reveal different failure modes: perceptual limitations, acoustic-cue reliance, and knowledge-based failures.

Question	Gold Answer	Transcript	Model Answer
Name the work.	Jurassic Park	Girl ouch Look at me Thanks for watching	"The transcription provided does not contain enough information to identify a specific work. It appears to be a mix of different languages and phrases, but without additional context or recognizable content, it's not possible to determine the name of a work based solely on this transcription."
What country is this national anthem from?	Armenia	You	"I'm sorry, but the transcription provided does not contain enough information to determine which country's national anthem it is. Could you provide more details or context from the transcription?"
Name the opera.	Madama Butterfly	(empty)	(no response)

Table 19: These examples illustrate that transcript-only inputs often fail when the audio contains non-verbal or partially verbal content. Even when the model has access to textual output, crucial identifying cues such as melody, instrumentation, or anthem structure are lost in the transcript. Therefore, audio is essential for correct resolution in AUDITA, and transcript degradation alone can explain a substantial portion of failures.